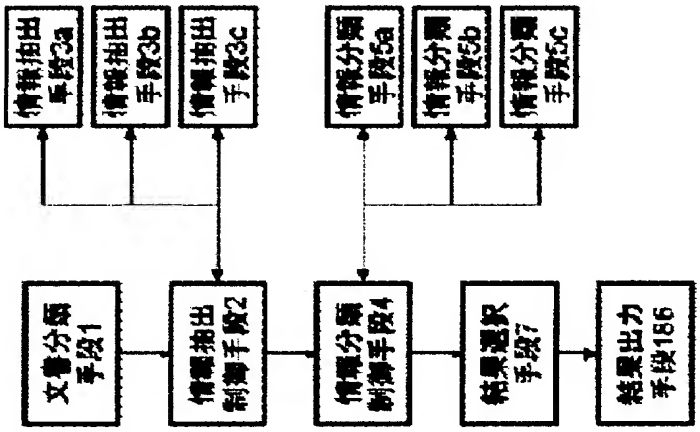
SYSTEM AND METHOD FOR INFORMATION EXTRACTION AND RECORDING MEDIUM STORED WITH RECORDED PROGRAM FOR INFORMATION EXTRACTION

Publication number: JP2001134600
Publication date: 2001-05-18
Inventor: YAMADA HIROSHI
Applicant: NIPPON ELECTRIC CO
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- European:
Application number: JP19990317069 19991108
Priority number(s): JP19990317069 19991108

Abstract of **JP2001134600**

PROBLEM TO BE SOLVED: To accurately extract information matching a purpose from a set of documents having different contents and formats. SOLUTION: A document classifying means 1 classifies the document set into specified categories. An information extracting means 3 judges which information is extracted according to the classifications of the documents and extracts information from the documents. An information classifying means 5 classifies the extracted information. A result selecting means 7 selects only necessary information from the extracted and classified information. A result output means 186 divides and outputs the information according to the classification result of the information.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－134600
（P2001－134600A）
(43)公開日　平成13年５月18日(2001.5.18)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３２０Ａ	５Ｂ０７５
		15/40	３７０Ａ	
		15/401	３１０Ｄ	

審査請求　有　　請求項の数18　ＯＬ　（全 15 頁）

(21)出願番号　特願平11－317069

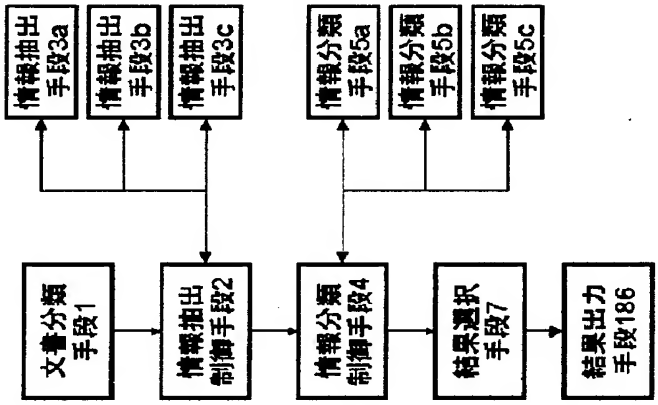
(22)出願日　平成11年11月８日(1999.11.8)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　山田　洋志
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100082935
弁理士　京本　直樹　（外２名）

Ｆターム(参考)　5B075 KK07 ND03 ND20 NR03 NR12 NS01 PQ02 UU06

(54)【発明の名称】　情報抽出システム、情報抽出方法および情報抽出用プログラムを記録した記録媒体

(57)【要約】
【課題】　内容や書式が異なる文書を含んだ文書集合から、目的に応じた情報を正確に抽出する。
【解決手段】　文書分類手段1で、文書集合を指定されたカテゴリーに分類する。情報抽出手段3で、文書の分類に応じてどの情報を抽出するかを判断し、文書から情報を抽出する。情報分類手段5で、抽出された各情報を分類する。結果選択手段7で、抽出・分類された情報のうち、必要な情報のみを選択する。結果出力手段186で、情報の分類結果に応じて情報を分割して出力する。



1

【特許請求の範囲】

【請求項１】入力された文書を分類する文書分類手段と、前記文書の分類に対応して該文書から抽出する情報とその抽出方法を変更する情報抽出手段と、前記文書の分類に応じ、前記情報抽出手段で抽出した情報を分類する情報抽出手段と、を少なくとも備えて構成されることを特徴とする情報抽出システム。

【請求項２】入力される文書集合に含まれる個々の文書を複数のカテゴリーに分類する文書分類手段と、
特定のカテゴリーに属する文書から抽出する情報の種類を定義した抽出情報定義手段と、
前記抽出情報定義手段を参照して該カテゴリーに分類された文書から前記抽出情報定義手段に定義されている情報を抽出する情報抽出手段と、
前記文書分類手段の文書分類結果に応じて適切な情報抽出手段を選択して文書から情報を抽出するための制御を行う情報抽出制御手段と、
特定のカテゴリーに属する文書から抽出された情報を分類する情報分類手段と、
該文書分類結果に応じて適切な情報分類手段を選択し、該文書から抽出された情報を分類するための制御を行う情報分類制御手段と、
前記文書分類手段による分類結果、前記情報抽出手段によって抽出された情報、前記情報分類手段による分類結果、を出力する結果出力手段とを備えたことを特徴とする情報抽出システム。

【請求項３】前記文書分類手段は、構造化された文書を対象とし、分類方法として文書のタイプを判別するタイプ判別手段を含むことを特徴とする請求項２記載の情報抽出システム。

【請求項４】前記文書分類手段は、特定のカテゴリーに属する文書を抽出する機能を有することを特徴とする請求項２または３記載の情報抽出システム。

【請求項５】前記結果出力手段は、特定の文書分類や情報分類に属する情報を選別して出力することを特徴とする請求項２から４のいずれか一項に記載の情報抽出システム。

【請求項６】前記結果出力手段は、前記情報抽出手段によって抽出した情報を、階層構造を持つ形式で出力することを特徴とする請求項２から５のいずれか一項に記載の情報抽出システム。

【請求項７】入力された文書を分類する文書分類ステップと、前記文書分類ステップで行われた前記文書の分類に対応して該文書から抽出する情報とその抽出方法を変更する情報抽出ステップと、前記文書分類ステップで行われた文書の分類に応じ、前記情報抽出ステップで抽出した情報を分類する情報分類ステップと、を少なくとも含むことを特徴とする情報抽出方法。

【請求項８】入力される文書集合に含まれる個々の文書を複数のカテゴリーに分類する文書分類ステップと、

2

特定のカテゴリーに属する文書から抽出する情報の種類を定義した抽出情報定義ステップと、
前記抽出情報定義ステップを参照して該カテゴリーに分類された文書から前記抽出情報定義ステップによって定義された情報を抽出する情報抽出ステップと、
前記文書分類ステップの文書分類結果に応じて適切な情報抽出ステップを選択して文書から情報を抽出するための制御を行う情報抽出制御ステップと、
特定のカテゴリーに属する文書から抽出された情報を分類する情報分類ステップと、
該文書分類結果に応じて適切な情報分類ステップを選択し、該文書から抽出された情報を分類するための制御を行う情報分類制御ステップと、
前記文書分類ステップによる分類結果、前記情報抽出ステップによって抽出された情報、前記情報分類ステップによる分類結果、を出力する結果出力ステップとを含むことを特徴とする情報抽出方法。

【請求項９】前記文書分類ステップは、構造化された文書を対象とし、分類方法として文書のタイプを判別するタイプ判別ステップを含むことを特徴とする請求項８記載の情報抽出方法。

【請求項１０】前記文書分類ステップは、特定のカテゴリーに属する文書を抽出する機能を有することを特徴とする請求項８または９記載の情報抽出方法。

【請求項１１】前記結果出力ステップは、特定の文書分類や情報分類に属する情報を選別して出力することを特徴とする請求項８から１０のいずれか一項に記載の情報抽出方法。

【請求項１２】前記結果出力ステップは、前記情報抽出ステップによって抽出した情報を、階層構造を持つ形式で出力することを特徴とする請求項８から１１のいずれか一項に記載の情報抽出方法。

【請求項１３】コンピュータに、
入力された文書を分類する文書分類ステップと、前記文書分類ステップで行われた前記文書の分類に対応して該文書から抽出する情報とその抽出方法を変更する情報抽出ステップと、前記文書分類ステップで行われた文書の分類に応じ、前記情報抽出ステップで抽出した情報を分類する情報分類ステップと、を少なくとも実行させることを特徴とする情報抽出用プログラムを記録した記録媒体。

【請求項１４】コンピュータに、
入力される文書集合に含まれる個々の文書を複数のカテゴリーに分類する文書分類ステップと、
特定のカテゴリーに属する文書から抽出する情報の種類を定義した抽出情報定義ステップと、
前記抽出情報定義ステップを参照して該カテゴリーに分類された文書から前記抽出情報定義ステップによって定義された情報を抽出する情報抽出ステップと、
前記文書分類ステップの文書分類結果に応じて適切な情

報抽出ステップを選択して文書から情報を抽出するための制御を行う情報抽出制御ステップと、
特定のカテゴリーに属する文書から抽出された情報を分類する情報分類ステップと、
該文書分類結果に応じて適切な情報分類ステップを選択し、該文書から抽出された情報を分類するための制御を行う情報分類制御ステップと、
前記文書分類ステップによる分類結果、前記情報抽出ステップによって抽出された情報、前記情報分類ステップによる分類結果、を出力する結果出力ステップと、
実行させることを特徴とする情報抽出用プログラムを記録した記録媒体。

【請求項１５】前記文書分類ステップは、構造化された文書を対象とし、分類方法として文書のタイプを判別するタイプ判別ステップを含むことを特徴とする請求項１４記載の情報抽出用プログラムを記録した記録媒体。

【請求項１６】前記文書分類ステップは、特定のカテゴリーに属する文書を抽出する機能を有することを特徴とする請求項１４または１５記載の情報抽出用プログラムを記録した記録媒体。

【請求項１７】前記結果出力ステップは、特定の文書分類や情報分類に属する情報を選別して出力することを特徴とする請求項１４から１６のいずれか一項に記載の情報抽出用プログラムを記録した記録媒体。

【請求項１８】前記結果出力ステップは、前記情報抽出ステップによって抽出した情報を、階層構造を持つ形式で出力することを特徴とする請求項１４から１７のいずれか一項に記載の情報抽出用プログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は文書中から特定の情報を抽出する情報抽出システム、情報抽出方法および情報抽出用プログラムを記録した記録媒体（以下、情報抽出システムと記載する。）に関し、特に文書を分類し、この分類結果に応じて抽出する情報を変えることができる情報抽出システムに関する。

【０００２】

【従来の技術】従来、この種の情報抽出システムの一例が、特開平８－３２９１６５号公報に記載されている。この従来の情報抽出システムの動作のフローチャートを図２３に示す。以下では、この図２３を用いてこの従来技術の動作を説明する。

【０００３】処理対象となる文書が入力されると(4801～4803)、その文書から特定のパターンを持つ文字列を数値データとして抽出し(4804)、さらに、その数値データの前後に存在する一定規則に基づいた文字列を数字文字列データとして抽出する(4805)。この数字文字列データの中から名詞データを抽出し、これを所定の項目毎に分類して、上記数値データを対応付ける(4806)。このようにして得られた各項目の数値データを１ヵ月等の所定期間毎に集計し(4807)、その集計結果データを表示する(4808)。

【０００４】この従来のシステムを利用すると家計簿のような特定形式の文書から、買物に使った金額を抽出することができる。

【０００５】次に、形式を問わない文書から情報を抽出する従来システムとしては、論文「固有名詞抽出システムの開発とIREX-NEにおける評価」(IREX-NEワークショップ予稿集、pp.171-178、1999)に記述されているシステムがある。このシステムでは、文書中から地名、人名、金額などを抽出できる。

【０００６】また、World Wide Web上で行われている情報サービスの１つに、就職情報やプレゼントの情報を集め、そこから会社名や賞品などの情報を人手で取り出したものをホームページ上でまとめて提示することが近年行われている。

【０００７】

【発明が解決しようとする課題】上記従来の技術の第１の問題点は、内容や形式にばらつきのある文書集合からは十分な精度で情報を抽出することができないということである。その理由は、従来の情報抽出システムでは特定の形式の文書を前提としているためである。

【０００８】次に第２の問題点は、必要のない情報も抽出してしまうということである。その理由は、どのような情報が必要かはそれぞれの文書の作成目的や利用目的によって変わってくるため、たとえば、金額や日付のような同じ情報であっても、出現する文書の種類や文書中の前後関係によって必要かどうかが変わってくるためである。 そのため、全ての文書、あるいは、文書全体から情報を抽出すると必要のない情報が含まれてしまう。

【０００９】次に第３の問題点は、第１、第２の問題点を補うために人手による作業を導入すると大量の文書に対応することができない、あるいは、非常に多くの時間や費用がかかるということである。

【００１０】よって本発明の目的は、上記従来技術の問題点を鑑み、不定形の文書集合からユーザの使用目的に応じた情報を抽出できる情報抽出システムを提供することにある。

【００１１】また本発明の他の目的は、必要な文書を容易に選択するための情報抽出システムを提供することにある。

【００１２】

【課題を解決するための手段】本発明の情報抽出システムは、文書を分類する文書分類手段と、文書の分類に対応して抽出する情報と抽出方法を変えることのできる情報抽出手段とを備え、文書の分類結果に応じた情報を抽出するよう動作する。

【００１３】さらに、文書の分類に応じて抽出した情報を分類することのできる情報抽出手段を備える。

【００１４】このような構成を採用し、各文書から必要な情報だけを正確に抽出することにより上記本発明の目的を達成することができる。

【００１５】

【発明の実施の形態】次に、本発明の実施の形態について図面を参照して詳細に説明する。

【００１６】（第１の実施の形態）

［構成の説明］図１を参照すると、本発明の第１の実施の形態は、文書分類手段1と、情報抽出制御手段2と、複数の情報抽出手段3と、情報分類制御手段4と、複数の情報分類手段5と、結果出力手段6から構成されている。

【００１７】情報抽出手段3は、文書の分類の種類に応じて複数存在し、抽出実行手段31と、抽出情報定義手段32と、抽出知識格納手段33とを含む。

【００１８】情報分類手段5は、文書の分類の種類に応じて複数存在し、分類実行手段51と、分類知識格納手段52とを含む。

【００１９】これらの手段はそれぞれ概略以下のように動作する。

【００２０】文書分類手段1は、複数の文書を入力とし、文書を指定されたカテゴリーに分類する。

【００２１】情報抽出制御手段2は、文書分類手段1の文書の分類結果に応じて使用する情報抽出手段3を決定し、決定した情報抽出手段3に文書から情報を抽出させる。また、情報抽出制御手段2よって動作される情報抽出手段3は、情報抽出制御手段2に指定された文書から情報を抽出する。

【００２２】情報分類制御手段4は、文書の分類結果に応じて使用する情報分類手段5を決定し、決定した情報分類手段５に文書から抽出した情報を分類させる。また、情報分類制御手段4によって動作される情報分類手段5は、情報分類制御手段4に指定された情報を分類する。

【００２３】結果出力手段6は、文書の分類結果、抽出された情報、情報の分類結果を出力する。

【００２４】ここで、情報抽出制御手段2と情報抽出手段3について、図２を参照して詳細に説明する。

【００２５】情報抽出手段3は、文書の分類カテゴリーの数に合わせて複数用意される。図２では3個の情報抽出手段3a、3b、3cが記述されているが、これは、情報抽出手段を3個に限定するものではない。

【００２６】情報抽出制御手段2は、文書分類手段1の分類結果に対応して情報抽出手段3を選択する。

【００２７】情報抽出手段3は、抽出実行手段31と、抽出情報定義手段32と、抽出知識格納手段33から構成されている。

【００２８】抽出実行手段31は、抽出情報定義手段32と抽出知識格納手段33を参照して、文書のカテゴリーに対応した情報を文書から抽出する。抽出情報定義手段32は、文書の分類カテゴリーに対応する情報の種類を格納する。抽出知識格納手段33は、抽出情報定義手段32に格納されている各情報を判別するための方法を格納する。

【００２９】情報分類制御手段4と情報分類手段5について図３を参照して説明する。

【００３０】情報分類手段5は、文書の分類カテゴリーの数に合わせて複数用意される。図３では3個の情報分類手段5a、5b、5cが記述されているがこれは、情報分類手段を3個に限定するものではない。

【００３１】情報分類手段5は、分類実行手段51と、分類知識格納手段52から構成されている。

【００３２】分類実行手段51は、分類知識格納手段52を参照して、情報抽出手段3で抽出した情報を分類する。分類知識格納手段52は、抽出した情報の種類ごとに情報を分類するための規則を格納している。

【００３３】［動作の説明］次に、図１及び図４のフローチャートを参照して本実施の形態の全体の動作について詳細に説明する。

【００３４】まず、文書分類手段1により文書をカテゴリーに分類する（図４のステップ401）。分類の方法としては、本出願人が本発明の出願前に出願したWebページの特徴を利用したタイプ分類技術（特願平10-200171号）や、単語情報を用いた分類技術を利用可能であるが、この文書分類手段1の構成を限定するものではない。文書分類手段1の構成としては、様々な既存技術を用いることができ、当業者であれば十分に理解し得るものである。また、一つの文書が複数のカテゴリーに分類される、あるいは、どこにも分類されない場合があってもよい。

【００３５】次に、情報抽出制御手段2によって、ステップ401における文書の分類結果に対応した情報抽出手段3を選択し、文書を分類させる（ステップ402）。

【００３６】情報抽出手段3は、文書から情報を抽出する（ステップ403）。各情報抽出手段3は、対応する文書カテゴリーに応じた情報を文書から抽出する。抽出実行手段31は、抽出情報定義手段32を参照して、文書から抽出する情報を得る。さらに、抽出実行手段31は、抽出知識格納手段33を参照して抽出する各情報を判別するための方法を得る。この判別のための方法としては、

・抽出する情報そのものの形式
・表や箇条書きなどにおける項目名や記述位置
・抽出する情報の前後に共起する単語や記号などの表記方法
・直接記述されていない情報を推定するための規則などがあり、さらに複数の方法を組み合わせることもできる。

【００３７】次に、情報分類制御手段4によって、文書の分類結果に対応した情報分類手段5を選択し、文書から抽出した情報を分類させる（ステップ404）。

【００３８】情報分類手段5は、対応している文書カテゴリーに応じて、抽出した情報を分類する（ステップ40

5)。分類実行手段51は、分類知識格納手段52を参照して、抽出した情報の種類ごとに情報を分類する。情報の分類知識としては、
・数値や時間の情報をいくつかの範囲に分割する
・単語と分類の対応表を用意する
・階層構造を持つ辞書を用いる
・文字列のパタンマッチによる分類などがあり、さらに複数の方法を組み合わせることもできる。

【００３９】最後に、結果を出力する(ステップ406)。文書名と抽出した情報の分類結果を出力する。また、文書カテゴリー、抽出した情報を出力するようにしてもよい。

【００４０】次に、本実施の形態の効果について説明する。

【００４１】本実施の形態では、文書を分類してから分類結果に応じて必要な情報を抽出するというように構成されているため、さまざまな種類の文書が混在している場合でも必要な情報だけを抽出できる。

【００４２】また、本実施の形態では、さらに、文書の分類結果に応じて情報を抽出する方法を選択するように構成されているため、高精度の情報抽出が可能になる。

【００４３】さらに、抽出した情報をもとに文書を参照することで、ユーザが必要な文書を容易に発見することが可能になる。

【００４４】[実施例]次に、具体的な実施例を用いて本実施の形態の動作を説明する。

【００４５】本実施例では、Webページ(HTMLファイル)から情報を抽出する場合を例にあげて説明する。

【００４６】分類対象となるWebページは、あらかじめ、自動収集プログラムやダウンロードプログラムによって記憶装置上に保存しておく。あるいは、WebページのURLの一覧を用意して、必要に応じてダウンロードするように構成してもよい。

【００４７】最初に、文書分類手段1によってWebページを分類する。

【００４８】本実施例では、文書分類手段1として、特願平10-200171号に記述されている構造化文書検索システムを利用する例を挙げる。ただし、先にも記載したように、本発明の文書分類手段1は、この例だけに限定されるものではない。このシステムでは、文書に対していくつかのタイプを設定し、各文書と該タイプの適合度を計算する。適合度の基準値を設定して、文書を基準値以上の適合度が得られたタイプに分類することができる。

【００４９】図５はWebページに対して適合度を計算した例を示す図である。図５で、“x.html”、“y.html”、“z.html”はページ名である。URLにはドメイン名やディレクトリ名が付くが図では省略した。また、「求人情報」、「イベント」、「プレゼント」が分類結果となるタイプ名で、数値が適合度であり、数値が高ければ高いほど、当該文書に含まれている内容がそのタイプと適合していることになる。ここで、この例における適合度の基準値を70とすると、“x.html”は「イベント」、“y.html”は「求人情報」、“z.html”は「プレゼント」に分類される。

【００５０】次に、情報抽出制御手段2によって各ページの分類結果に対応する情報抽出手段3を選択する。「求人情報」、「イベント」、「プレゼント」に分類されたページからは、それぞれ、情報抽出手段3a、3b、3cによって情報を抽出する。

【００５１】図６は、情報抽出手段3のそれぞれの情報名定義手段32が格納する情報の種類の例を示す図である。情報抽出手段3a、3b、3cの抽出情報定義手段32a、32b、33c、は、それぞれ図６の6a、6b、6cに示す情報を格納している。例えば、「求人情報」に分類される文書から情報を抽出する役目を情報抽出手段3aが持っているときには、抽出情報定義手段32aには、図６の6aに示す情報をもち、抽出実行手段31aは、この抽出情報定義手段32aの定義（この例では、[勤務地]、「職種」）の内容に沿った情報を当該文書から抽出する。また、この例の場合、情報抽出制御手段2には、複数の情報抽出手段がそれぞれどの分類の情報を担当するかを判断する情報が必要となる。

【００５２】図７は、抽出知識格納手段33が格納する情報抽出の方法の例を示す図である。図７の(a)、(b)、(c)はそれぞれ抽出知識格納手段33a、33b、33cに対応する。

【００５３】図７では、図６に示した抽出情報定義手段32の各タイプの「情報名」ごとに、どのような方法で情報を見つけるかを記述している。図７中で、「記述種類」は、どこに書かれている情報を抽出するかを指定している。例えば、「見出し」であれば表や箇条書きの中で「パタン」に書かれている見出しが付いている部分を抽出する。また、「指定タグ」であれば特定のタグ内を抽出し、さらに「テキスト」であればテキスト中で「パタン」に一致する部分を抽出する。

【００５４】また、図７中のパタンで%dや%sとあるのは変数の意味で、任意の文字列が入ることを表す。数値の範囲や文字の種類や長さなど、詳細な指定を加えられる書式にすることで、より正確に抽出にすることができる。複数の抽出方法がある場合は、「優先度」によって順序づけしている。

【００５５】図７の例で、「求人情報」タイプの「勤務地」と「イベント」タイプの「開催地」はいずれも場所に関する情報であるが、異なるパタンを利用して抽出が行われるため、区別して抽出できる。同様に「イベント」タイプの「開催日」と「プレゼント」タイプの「応募〆切」も日付であるが区別して抽出できる。

【００５６】よって、図５の例で「求人情報」に分類された“y.html”については、図６の6aから、「勤務地」と「職種」を抽出し、抽出の際には、図７の7aの抽出方

法を参照して抽出することになる。

【００５７】図８は、「求人情報」に分類されるWebページの表示イメージの例である。実際のファイルでは、タグによって表や箇条書きを表現している。ここから情報を抽出する場合について説明する。

【００５８】「勤務地」については、図７(a)によって、表や箇条書きの中から「勤務地」あるいは「勤務場所」、「営業所」という見出しのある項目を探し、「川崎市」を抽出する。同様に「職種」の項から「システムエンジニア」を抽出する。

【００５９】図５の例で「イベント」に分類された“x.html”については、図６の6bから、「名称」と「開催地」、「開催日」を抽出する。

【００６０】図９は、「イベント」のWebページの表示イメージの例である。このページから箇条書きの見出しを元にして、「名称」と「開催地」を抽出する。「開催日」は箇条書きによる記述がないので、図７(b)からテキスト中からパタンに一致する文字列を探し、「1999年10月10日」を抽出する。

【００６１】図１０は、各ページから抽出した情報の例を示す図である。ページ名、分類結果、情報名、抽出した内容が記述されている。

【００６２】なお、抽出情報定義手段32で定義されているすべての情報が抽出できるとは限らない。たとえば、「イベント」に分類された文書から「名称」と「開催日」が抽出され「開催地」は抽出されない場合もありうる。

【００６３】次に、情報分類制御手段4によって各ページの分類結果に対応する情報分類手段5を選択する。「求人情報」、「イベント」、「プレゼント」に分類されたページから抽出された情報は、それぞれ、情報分類手段5a、5b、5cによって分類する。

【００６４】図１１は、分類知識格納手段52に記述する分類方法の概要を示す図である。各情報によって分類する種類と、方法を記述する。図１１(a)(b)(c)はそれぞれ分類知識格納手段53a、53b、53cに記述する。

【００６５】図１２は、分類結果を表す図である。図１２で「イベント」の開催日の分類は6桁の数字で年と月で表している。「プレゼントの」〆切日は単純に分類しているので分類に対応する週の日曜日の日付で表している。

【００６６】分類実行手段51は、分類知識格納手段52の記述に従って情報を分類する。分類の方法としては、

・各分類に含まれる単語のリストを用いる
・都道府県ごとに市町村名や施設名を記述したシソーラスを用いる
・単語シソーラスを用いて上位概念にまとめるなどの方法がある。

【００６７】情報の種類ごとに分類方法を指定することで、図１１の「イベント」タイプの「開催日」と「プレゼント」タイプの「応募〆切」のように、同じ日付の情報であっても分類を変えることもできる。

【００６８】最後に、結果出力手段6によって、抽出した情報と分類結果を出力する。

【００６９】出力方法としては、抽出した情報をCVSファイル形式など、検索システムやデータベースシステムなどのシステムに登録できる形式で記憶装置上に出力するほか、図１２のような一覧表やHTMLなどの表示できる形式、その他XML、SGMLなどの構造化した文書形式が使用できる。

【００７０】なお、本実施例は、HTMLに限らず、SGMLやXMLなど構造化された文書に対して同様に機能する。

【００７１】［別の実施例］次に、別の実施例を用いて本実施の形態の動作を説明する。

【００７２】ここでは、新聞記事から情報を抽出する場合を説明する。

【００７３】まず、新聞記事をあらかじめ定めたカテゴリーに分類する。記事中に含まれる単語を元に文書を分類する従来システム(たとえばジャストシステム社のCB Classifier（商標）など)が利用できる。たとえば、記事を「国際政治」、「新製品情報」、「スポーツ」に分類する。

【００７４】次に、情報抽出制御手段2によって各記事の分類に対応する情報抽出手段3を選択し情報を抽出する。

【００７５】図１３は、情報名定義手段32が格納する情報の種類の例を示す図である。たとえば、「国際政治」に分類されたページからは「地名」と「関係者」の情報を抽出する(図１３の13a)。

【００７６】図１４は、抽出知識格納手段23が格納する、情報抽出の方法の例を示す図である。新聞記事が構造化されていないテキストであるので、テキスト中でのパタンマッチで情報を抽出する。「国際政治」の記事から抽出する「関係者」については肩書きを用いて国の代表者レベルの人物に限定している（図１４(a)）。

【００７７】図１４(a)、(c)では、「地名」と「競技名」については、あらかじめリストを作成してテキストを探索する。

【００７８】図１５は、新聞記事から抽出した情報の例を示す図である。記事番号、分類結果、情報名、抽出した内容が記述されている。

【００７９】次に、情報分類制御手段4によって、記事の分類結果に対応した情報分類手段5を選択し、記事から抽出した情報を分類する。図１６は、分類知識格納手段52に記述する分類方法の概要を示す図である。各情報によって分類する種類と、方法を記述する。図１７は、分類結果を表す図である。

【００８０】最後に、結果出力手段6によって、抽出した情報と分類結果を出力する。

【００８１】（第２の実施の形態）次に、本発明の第２

の実施の形態について図面を参照して詳細に説明する。

【００８２】図１８を参照すると、本発明の第２の実施の形態は、文書分類手段1と、情報抽出制御手段2と、複数の情報抽出手段3と、情報分類制御手段4と、複数の情報分類手段5と、結果選択手段7と、結果出力手段6から構成されている。

【００８３】次に、図１８および図１９のフローチャートを参照して本実施の形態の全体の動作について詳細に説明する。

【００８４】文書分類手段1、情報抽出制御手段2、情報抽出手段3、情報分類制御手段4、情報分類手段5の動作(ステップ1901～1905)は第１の形態と同じである。

【００８５】結果選択手段7は、抽出・分類された情報のうち、特定の情報のみを選択して結果出力手段186に渡す(ステップ1906)。選択基準としては、

・文書の分類を指定する
・情報の分類を指定する
・抽出した情報に条件を指定する
・特定の情報が抽出できた文書からの情報のみを選択するなどがある。

【００８６】結果出力手段186は、結果選択手段5の出力を受け取り、情報の分類結果に応じて複数に分割して出力する(ステップ1907)。分割方法としては、

・文書の分類ごとに分割する
・情報の分類ごとに分割する
・複数の分類の組み合わせごとに分割するなどがある。

【００８７】次に、本実施の形態の効果について説明する。

【００８８】本実施の形態では、抽出あるいは分類した情報から結果選択手段7によって、必要なものだけを選択して出力するというように構成されているため、特定の目的やユーザにあわせた情報抽出結果を提供できる。

【００８９】また、本実施の形態では、さらに、結果出力手段186によって、抽出した情報を分類して分割して出力するというように構成されている。文書の分類と抽出した情報の分類を組み合わせることで、特定の情報を持つ文書を他と区別することができ、特定の目的を持って文書を探す場合に容易に目的を達成できる。

【００９０】［実施例］次に、具体的な実施例を用いて本実施の形態の動作を説明する。

【００９１】ここでは、Webページ(HTMLファイル)から抽出した情報を例に説明する。

【００９２】情報分類制御手段4および情報分類手段5による分類結果として図７に示す形式の情報が得られる。

【００９３】次に、結果選択手段5は結果から特定の情報を選択する。選択方法の例として、

・「イベント」ページから抽出された情報を選択する
・「勤務地」が関東である「求人情報」から抽出された情報を選択する
・「開催日」が現在の日時より先である「イベント」ページから抽出された情報を選択する
・「賞品」と「応募締め切り」の両方が抽出できた「プレゼント」ページから抽出された情報を選択するなどがある。

【００９４】最後に、結果出力手段184は、結果選択手段5の出力を受け取り、情報の分類結果に応じて複数に分割して出力する。分割方法としては、

・ページの分類ごとに分割する
・「イベント」ページの情報を「開催地」の都道府県別に分割する
・「イベント」ページを「開催日」の月別に分割する
・「求人情報」ページの情報を、「勤務地」の都道府県別分類と「職種」の分類組み合わせで分割する。この場合、（47都道府県×職種分類の数）通りに分割することになるなどがある。

【００９５】さらに、ページの分類と抽出された情報の分類を階層的に組み合わせた出力形式を使用することで、特定の情報を含んでいる文書を効率よく見つけ出すことができる。階層構造の実現方法としては、たとえば、HTMLやXMLのハイパーテキストの機能を使うことで実現できる。

【００９６】図２０は、結果出力手段186の出力結果の例を示す図である。この例では、文書の分類結果の一覧(2001)から、各分類の文書から抽出された情報の分類にリンクがはられている。さらに、情報の分類から個々の文書名の一覧を参照できる。

【００９７】たとえば、「求人情報」からのリンクの一つとして「勤務地」と「職種」の組み合わせの一覧(2002)がリンクし、さらに該当する求人情報の一覧(2003)にリンクしている。「イベント」については、「開催日」による分類(2004)からイベント一覧(2005)にリンクしている。

【００９８】このような出力形式を用いることで、出力結果から目的とする文書を見つけ出すことが容易になる。

【００９９】（第３の実施の形態）次に、本発明の第３の実施の形態について図面を参照して詳細に説明する。

【０１００】図２１を参照すると、本発明の第３の実施の形態は、文書分類手段2101と、情報抽出制御手段2と、情報抽出手段3と、情報分類制御手段4と、情報分類手段5と、結果選択手段7と、結果出力手段6から構成されている。

【０１０１】次に、図１８を参照して本実施の形態の全体の動作について詳細に説明するが、情報抽出制御手段2、情報抽出手段3、情報分類制御手段4、情報分類手段5、情報出力手段6の動作は第１の実施の形態と同じであるため説明を省略する。

【０１０２】本実施の形態の情報分類手段2101は、文書を分類して特定のカテゴリーに含まれる文書のみを情報抽出制御手段2に渡す。

【０１０３】次に、本実施の形態の効果について説明する。

【０１０４】本実施の形態では、特定のカテゴリーに分類される文書のみを対象にして情報抽出処理を行うため、必要な情報を効率よく抽出することができる。

【０１０５】なお、本実施の形態の特殊な場合として、特定のカテゴリーが１種類だけの場合は、カテゴリーによる情報抽出手段、情報分類手段の選択が不要になるため情報抽出制御手段と情報分類制御手段を省略して構成できる(図２２)。

【０１０６】また、本実施の形態における情報抽出装置をコンピュータによって実現するには、第１の実施の形態であれば、文書分類手段１、情報抽出制御手段２、情報抽出手段３、情報分類制御手段４、情報分類手段５、結果出力手段６の各機能を実現するコンピュータプログラムを作成し、そのコンピュータプログラムをＣＤ－ＲＯＭやフロッピー（登録商標）ディスクや半導体メモリに代表される記録媒体に記録しておき、コンピュータ側では、このプログラムが記録された記録媒体を読み出すことにより、コンピュータに上記各機能を生成すれば、本発明の情報検索装置をコンピュータによって実現することができる。また、このコンピュータプログラムは、例えばサーバ内の記録装置に記録されている形態でもかまわなく、ネットワークを介してこのサーバ内に含まれるプログラムを提供する形態でもよい。

【０１０７】

【発明の効果】本発明の第１の効果は、情報を正確に抽出できることにある。その理由は、文書を分類して、分類結果に応じた情報抽出方法を定義するためである。

【０１０８】また、本発明の第２の効果は、必要な情報だけを抽出できることにある。その理由は、文書を分類して分類に応じて重要な情報だけを抽出するためである。

【０１０９】さらに、本発明の第３の効果は、大量の文書から情報を抽出できることにある。その理由は、正確に情報を抽出できるため人手を使わず自動的に情報を抽出できるためである。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態の構成を示すブロック図である。

【図２】第１の実施の形態の情報抽出手段3の構成を示すブロック図である。

【図３】第１の実施の形態の情報分類手段5の構成を示すブロック図である。

【図４】第１の実施の形態の動作を示す流れ図である。

【図５】第１の実施例でWebページに対して適合度を計算した例を示す図である。

【図６】第１の実施例の抽出情報定義手段32が格納する情報の種類の例を示す図である。

【図７】第１の実施例の抽出知識格納手段33が格納する、情報抽出の方法の例を示す図である。

【図８】第１の実施例で「求人情報」のWebページの表示イメージの例である。

【図９】第１の実施例で「イベント」のWebページの表示イメージの例である。

【図１０】第１の実施例の情報抽出手段3が抽出した情報の例を示す図である。

【図１１】第１の実施例の分類知識格納手段52に記述する分類方法の概要を示す図である。

【図１２】第１の実施例の分類結果を表す図である。

【図１３】第２の実施例の抽出情報定義手段32が格納する情報の種類の例を示す図である。

【図１４】第２の実施例の抽出知識格納手段33が格納する、情報抽出の方法の例を示す図である。

【図１５】第２の実施例の情報抽出手段3が抽出した情報の例を示す図である。

【図１６】第２の実施例の分類知識格納手段52に記述する分類方法の概要を示す図である。

【図１７】第２の実施例の分類結果を表す図である。

【図１８】本発明の第２の実施の形態の構成を示すブロック図である。

【図１９】第２の実施の形態の動作を示す流れ図である。

【図２０】第２の実施の形態の実施例の出力手段1806の出力例を示す図である。

【図２１】本発明の第３の実施の形態の構成を示すブロック図である。

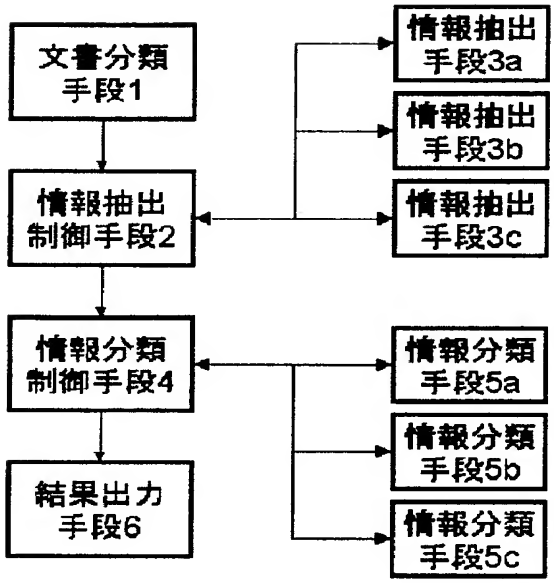
【図２２】第３の実施の形態の別の構成を示すブロック図である。

【図２３】従来の文書検索装置の動作を示す流れ図である。

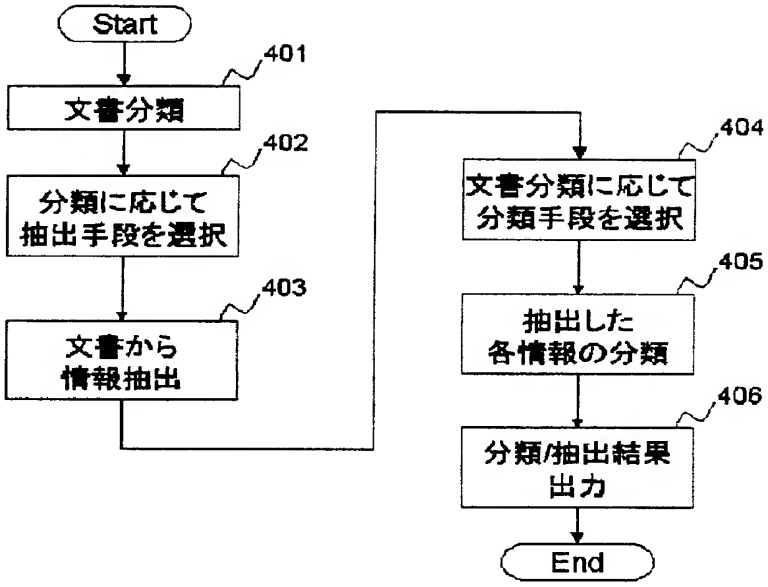
【符号の説明】

1、2101　文書分類手段
2　情報抽出制御手段
3　情報抽出手段
31　抽出実行手段
32　抽出情報定義手段
33　抽出知識格納手段
4　情報分類制御手段
5　情報分類手段
51　分類実行手段
52　分類知識格納手段
6、186　結果出力手段
7　結果選択手段

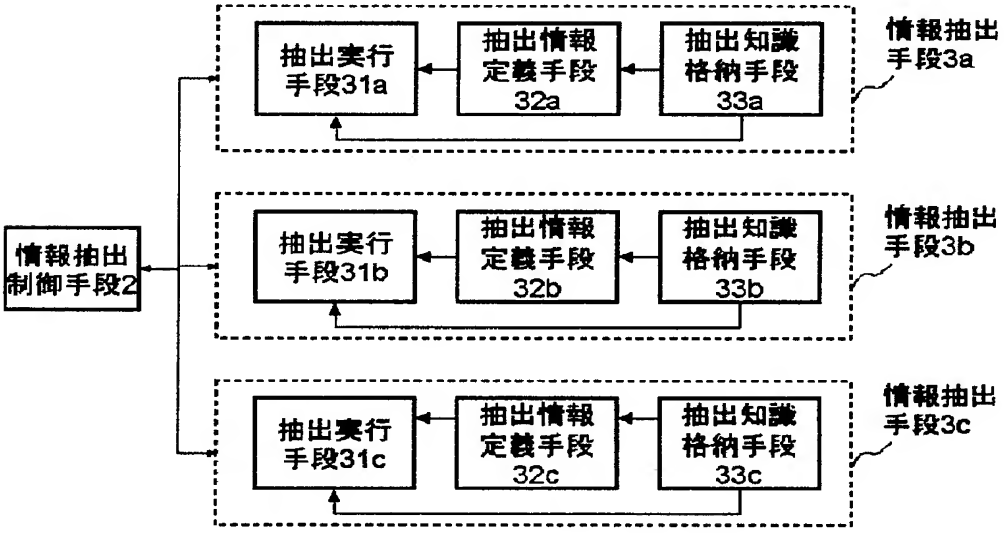
【図１】



【図４】



【図２】



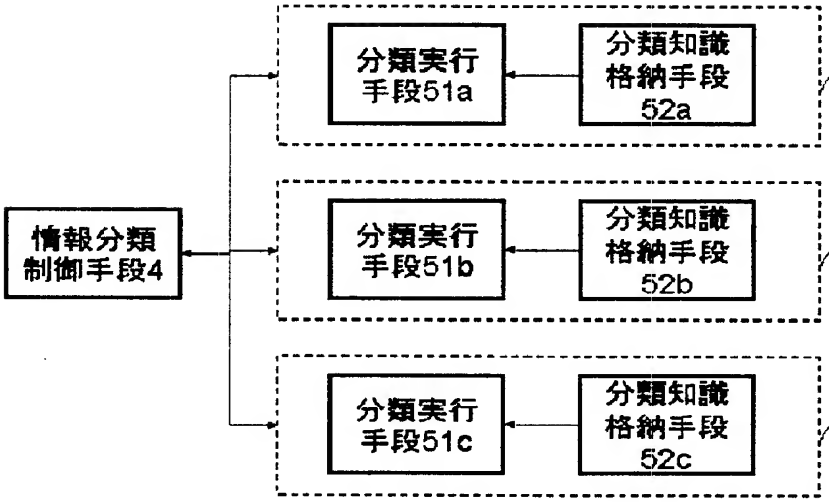
【図５】

	求人情報	イベント	プレゼント
x.html	30	100	24
y.html	80	0	6
z.html	10	25	85

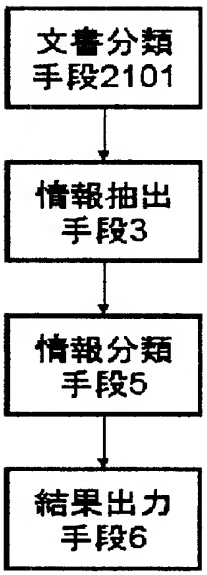
【図６】

タイプ	情報の種類	
求人情報	勤務地,職種	6a
イベント	名称,開催地,開催日	6b
プレゼント	賞品,応募〆切	6c

【図３】



【図２２】



【図７】

(a)

タイプ	情報名	記述種類	優先度	パタン
求人情報	勤務地	見出し	1	勤務地,勤務場所,営業所
求人情報	職種	見出し	1	職種,業務内容

(b)

タイプ	情報名	記述種類	優先度	パタン
イベント	名称	見出し	1	イベント,名称
イベント	名称	指定タグ	2	<TITLE>
イベント	開催地	見出し	1	会場,開催地,開催場所
イベント	開催日	見出し	1	開催日,日時
イベント	開催日	テキスト	2	%d 年%d 月%d 日, %d/%d/%d,%d 年%d 月

(c)

タイプ	情報名	記述種類	優先度	パタン
プレゼント	賞品	見出し	1	賞品,プレゼント
プレゼント	賞品	テキスト	2	%s が当たる
プレゼント	応募〆切	見出し	1	〆切,締め切り

【図９】

1999年10月10日の行事

・イベント: ○○区運動会
・会場: 札幌市立○○小学校グラウンド

【図１３】

分類	情報の種類	
国際政治	地名,関係者名	13a
新製品	製品名,価格	13b
スポーツ	競技名,選手,開催地	13c

【図８】

(a)

会社名	○○システム(株)
職種	システムエンジニア
勤務地	川崎市
給与	当社規定による

(b)

- 会社名: ○○システム(株)
- 職種: システムエンジニア
- 勤務地: 川崎市
- 給与: 当社規定による

【図１０】

ページ	タイプ	情報名	情報
x.html	求人情報	勤務地	川崎市
		職種	システムエンジニア
z.html	イベント	名称	○○区運動会
		開催地	札幌市立○○小学校グラウンド
		開催日	1999年10月10日
y.html	プレゼント	賞品	新巻鮭
		応募〆切	12月14日
		応募方法	はがき

【図１２】

ページ	タイプ	情報名	情報	分類
x.html	求人情報	勤務地	川崎市	神奈川県
		職種	システムエンジニア	技術
z.html	イベント	名称	○○区運動会	
		開催地	札幌市立○○小学校グラウンド	北海道 札幌市
		開催日	1999年10月10日	199910
y.html	プレゼント	賞品	新巻鮭	食品
		応募〆切	12月14日	19991212
		応募方法	はがき	葉書

【図１１】

(a)

タイプ	情報名	分類方法の説明
求人情報	勤務地	都道府県別に分類． 地名シソーラスを使用．
	職種	技術・営業・事務・その他 に分類． 分類用リストを使用

(b)

タイプ	情報名	分類方法の説明
イベント	名称	分類せず
	開催地	市区まで分類． 地名シソーラスを使用．
	開催日	年月別に分類．

(c)

タイプ	情報名	分類方法の説明
プレゼント	賞品	カテゴリー別分類． 単語シソーラスを使用．
	応募〆切	週単位で分類
	応募方法	葉書・Ｅメール・フォーム・電話 に分類．分類用リストを使用．

【図１４】

(a)

分類	情報名	記述種類	優先度	パタン
国際政治	地名	テキスト	1	地名リストを参照
国際政治	関係者	テキスト	1	%s 国王,%s 大統領, %s 首相

(b)

分類	情報名	記述種類	優先度	パタン
新製品	製品名	テキスト	1	%s を発売,%s を発表
新製品	価格	テキスト	1	%d 円,%d ドル

(c)

分類	情報名	記述種類	優先度	パタン
スポーツ	競技名	テキスト	1	競技名リストを参照
スポーツ	開催地	テキスト	1	地名リストを参照
スポーツ	選手名	テキスト	1	%s 選手

【図１５】

記事番号	タイプ	情報名	情報
001	国際政治	地名	カリフォルニア州
		代表者	クリントン大統領
		代表者	小渕首相
002	新製品	製品名	PC2000/XX
		価格	25 万円
003	スポーツ	競技名	野球
		開催地	韓国
		選手名	松坂選手

【図１６】

(a)

タイプ	情報名	分類方法の説明
国際政治	地名	国別に分類． 都市シソーラス，国名リストを使用．
	関係者	分類せず

(b)

タイプ	情報名	分類方法の説明
新製品	製品名	製品分野で分類． 分野ごとのキーワードリスト使用
	価格	分類せず．

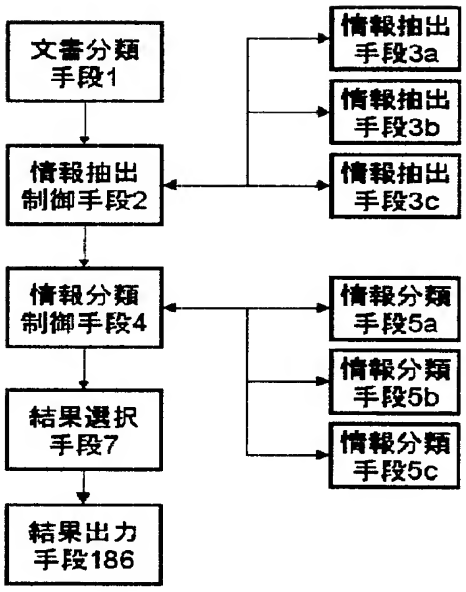
(c)

タイプ	情報名	分類方法の説明
スポーツ	競技名	分類せず
	選手名	分類せず
	開催地	国内と国外に分類． 国内地名シソーラス使用

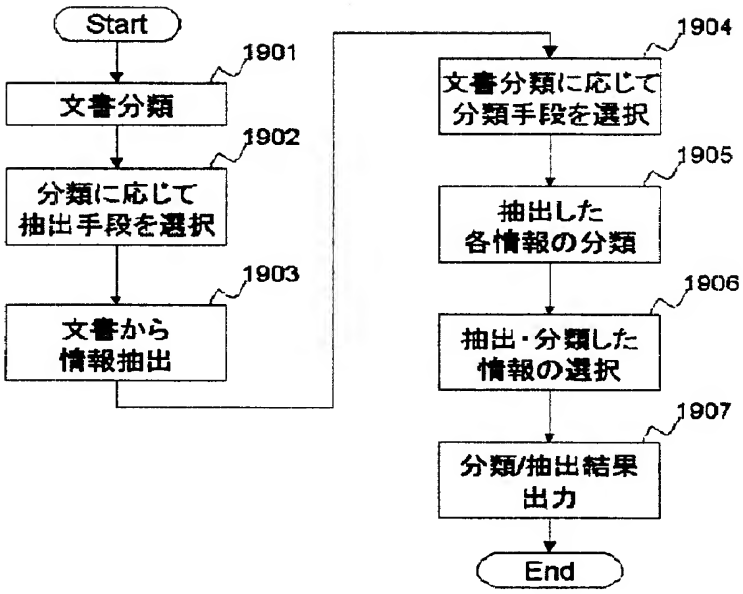
【図１７】

記事番号	タイプ	情報名	情報	分類
001	国際政治	地名	カリフォルニア州	USA
		代表者	クリントン大統領	
		代表者	小渕首相	
002	新製品	製品名	PC2000/XX	
		価格	25 万円	
003	スポーツ	競技名	野球	
		開催地	韓国	国外
		選手名	松坂選手	

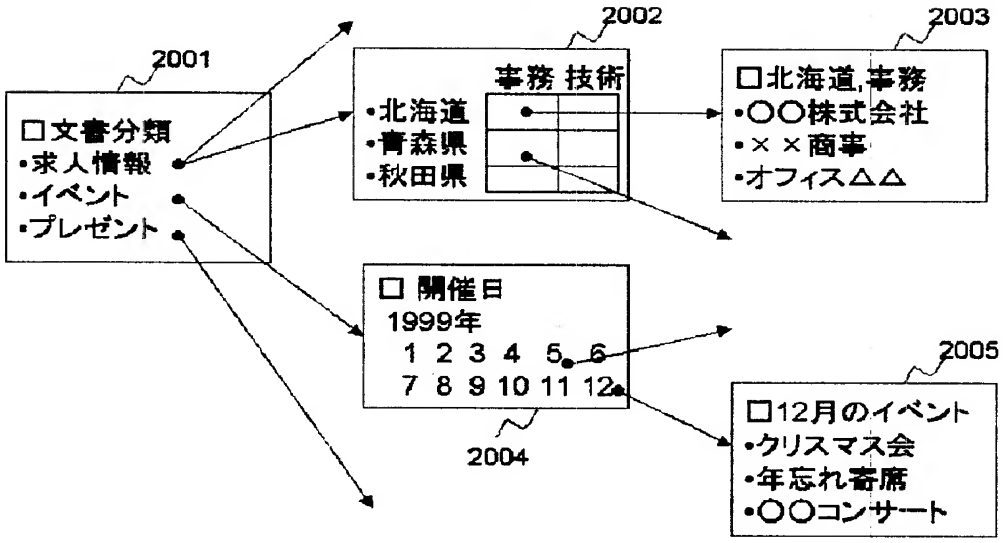
【図１８】


文書分類
手段1
情報抽出
制御手段2
情報分類
制御手段4
結果選択
手段7
結果出力
手段186
情報抽出
手段3a
情報抽出
手段3b
情報抽出
手段3c
情報分類
手段5a
情報分類
手段5b
情報分類
手段5c

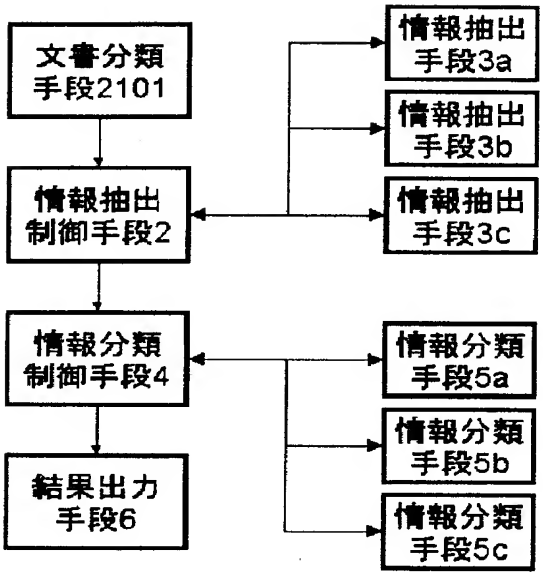
【図１９】


Start
1901
文書分類
1902
分類に応じて
抽出手段を選択
1903
文書から
情報抽出
1904
文書分類に応じて
分類手段を選択
1905
抽出した
各情報の分類
1906
抽出・分類した
情報の選択
1907
分類/抽出結果
出力
End

【図２０】


2001
□文書分類
•求人情報
•イベント
•プレゼント
2002
事務 技術
•北海道
•青森県
•秋田県
2003
□北海道,事務
•○○株式会社
•××商事
•オフィス△△
□ 開催日
1999年
1 2 3 4 5 6
7 8 9 10 11 12
2004
2005
□12月のイベント
•クリスマス会
•年忘れ寄席
•○○コンサート

【図２１】


文書分類
手段2101
情報抽出
制御手段2
情報分類
制御手段4
結果出力
手段6
情報抽出
手段3a
情報抽出
手段3b
情報抽出
手段3c
情報分類
手段5a
情報分類
手段5b
情報分類
手段5c

【図２３】


Start
4801
文書種類の選択
4802
文書ボタンをクリックして文書を開く
4803
文書入力
4804
種数値データ抽出、格納
4805
数字文字列データ抽出、格納
4806
名詞データ分類格納
4807
データ集計
4808
集計結果表示
End